视觉里程计算法研究综述 *

慈文彦 [1,2]，黄影平 [1†]，胡　兴 [1]

(1. 上海理工大学 光电信息与计算机工程学院，上海 200093; 2. 南京师范大学泰州学院 电力工程学院，江苏 泰州 225300)

摘　要：视觉里程计通过分析相机所获取的图像流信息估计移动机器人的位姿。为了深入分析视觉里程计算法的发展现状，结合一些先进的视觉里程计系统，综述了视觉里程计的相关技术以及最新的研究成果。首先简述了视觉里程计的概念和发展历程，介绍了视觉里程计问题的数学描述和分类方法；然后，详细阐述了视觉里程计的关键技术，包括特征模块、帧间位姿估计和减少漂移；此外，还介绍了基于深度学习的视觉里程计的发展动态。最后，总结了视觉里程计目前存在的问题，展望了未来的发展趋势。

关键词：机器视觉；视觉里程计；位姿估计；视觉导航；移动机器人；深度学习

中图分类号：TP242.6　　doi: 10.3969/j.issn.1001-3695.2018.05.0346

Review of visual odometry algorithms

Ci Wenyan[1, 2], Huang Yingping[1, †], Hu Xing[1]

(1. *School of Optical-Electrical & Computer Engineering*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*; 2. *School of Electric Power Engineering*, *Nanjing Normal University Taizhou Colledge*, *Taizhou Jiangsu* 225300, *China*)

Abstract: Visual odometry (VO) estimates the pose of a mobile robot by analyzing the image flow captured by the equipped cameras. In order to analyze the development of VO algorithms, this paper reviewed the related technologies of VO and the up-to-date research state combined with some advanced VO systems. Firstly, this paper described the concept and the evolution of VO, and introduced the mathematical description and the classification of VO. Then, it analyzed the key technologies of VO in details, including feature selection, motion estimation and drift reduction. In addition, it also introduced the latest deep learning based VO. At last, it discussed the existing problems and prospects the development trend of VO.

Key words: machine vision; visual odometry; pose estimation; vision-based navigation; mobile robots; deep learning

0　引言

在移动机器人系统中，要进行目标探测和定位，对于自身位姿的估计非常重要。传统的位姿估计方法有 GPS、IMU、轮速传感器和声纳定位系统等里程计技术。近年来，相机系统变得更加便宜，分辨率和帧率也更高，计算机的性能有了显著提高，实时的图像处理成为可能。一种新的位姿估计方法因此而产生，即视觉里程计（vision odometry,VO）。VO 仅利用单个或多个相机所获取的图像流估计智能体位姿[1]。它的成本较低，能够在水下和空中等 GPS 失效的环境中工作，其局部漂移率小于轮速传感器和低精度的 IMU，它所获得的数据能够很方便的和其他基于视觉的算法融合，省去了传感器之间的标定。

从连续的图像序列中估计相机自运动的思想最早由 Moravec 等人[2]提出。他利用一个可滑动的相机获取视觉信息完成了机器人的室内导航。1987 年，Matthies 等人[3]设计了从特征提取、特征匹配与跟踪到运动估计的理论框架，该框架至今仍为大多数 VO 系统所遵循。绝大多数早期的 VO 系统主要应用于行星探测[2, 4]，其中最典型的当属美国 NASA 的火星探测项目。VO 在火星探测器中用于在轮速传感器失效时测量 6 个自由度参数。VO 这个术语是由 Nister 等人[5]在 2004 年创造的。他们设计了一种实时的 VO 系统，真正意义上实现了机器人室外运动导航，同时还提出了两类 VO 的实现途径和流程，即单目视觉和立体视觉的方法，这为后来 VO 的研究奠定了新的基础。

与 VO 紧密相关的一个领域是基于视觉的同时定位与地图构建技术(visual simultaneous localization and mapping,V-SLAM)[6-8]。V-SLAM 在一个未知的环境中对机器人进行自定位并实时重建环境的三维结构。它的目标是获得对机器人运动轨

收稿日期：2018-05-28；**修回日期：**2018-07-14　**基金项目：**国家自然科学基金资助项目(61374197)；国家教育部博士点基金资助项目(20133120110006)；泰州市科技支撑计划（社会发展）资助项目（TS201701）

作者简介：慈文彦（1982-），男，黑龙江哈尔滨人，讲师，博士研究生，主要研究方向为计算机视觉、视觉里程计（wenyantz@163.com）；黄影平（1966-），男（通信作者），教授，博士，主要研究方向为智能汽车、模式识别、汽车电子；胡兴（1983-），男，讲师，博士，主要研究方向为机器学习、计算机视觉.

迹的全局一致性估计，这意味着机器人必须能够识别曾经到过的地方，这个过程被称为闭环检测。而 VO 是增量式的重构路径，它只关心轨迹的局部一致性。从实时性和环境适应性的角度出发，专注于局部运动估计的 VO 更具有实用价值，更适用于大范围运动的移动机器人。

过去曾经出现过一些有关 VO 的综述文献[1, 9-11],尤其是 Scaramuzza 等人[1, 11]的两篇文章系统的介绍了 VO 在 2012 年以前的发展状况。但是，近几年 VO 技术取得了很大的进步，随着大量高性能 VO 系统的涌现，这些文献已经不能反映最新的 VO 技术的发展。本文在综述过程中，侧重结合了一些先进的 VO 系统。文章首先介绍了 VO 的概况，包括 VO 问题的数学描述及其分类；然后重点综述了特征模块、帧间位姿估计和减少漂移等 VO 的关键技术。针对近几年来出现的基于深度学习的 VO，简述了它的发展动态并分析了它的优势和不足。考虑到算法评价对于 VO 发展的重要性，文章还介绍了三个常用的 VO 公共数据集。最后总结了 VO 目前存在的问题，展望了它的发展趋势。

1 视觉里程计概况

1.1 视觉里程计问题数学描述

摄像机模型是将三维世界投影到二维图像平面的函数。有很多摄像机模型，如透视投影模型、全方向摄像机模型、球形模型等，其中最基本也是最常用的是透视投影模型。在透视投影中，远的目标要比近的目标看起来小一些，这个性质与人类视觉以及大多数摄像机是相符的。透视投影的几何关系如图 1 所示。其中 C 点称为摄像机光心，由点 C 与 x 、 y 、 z 轴组成的直角坐标系称为摄像机坐标系。 x 、 y 轴与图像的 u 、 v 轴平行， z 轴为摄像机光轴，它与图像平面垂直。光轴与图像平面的交点 (c_u, c_v) 即为图像坐标系的原点。图中点 $P=[X,Y,Z]^T$ 为三维世界中的一点，点 $p=[u,v]^T$ 是它在图像平面上的投影。

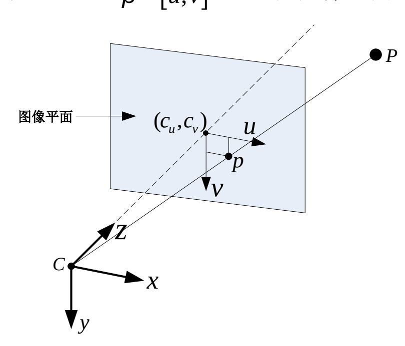


图 1 摄像机透视投影模型

由三维世界到二维图像平面的透视投影方程可以表示为如下形式：

$$Z\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f_u&0&c_u\\0&f_v&c_v\\0&0&1\end{bmatrix}\begin{bmatrix}X\\Y\\Z\end{bmatrix}=KP \tag{1}$$

这里： f_u、 f_v 分别为 x 、 y 方向的焦距， f_u、 f_v、 c_u、 c_v 只与摄像机内部结构有关，因此称为摄像机内部参数， K 称为摄像机内部参数矩阵。

假设智能体在环境中运动，装载在智能体上的相机与智能体之间不存在相对运动。相机在离散时间 k 拍摄图像，在各个时刻所拍摄的图像序列可以表示为 $I_{0:n}=\{I_0,\ldots,I_n\}$ 。从时刻 $k-1$ 到时刻 k 的坐标变换 $T_{k,k-1}\in\mathbb{R}^{4\times4}$ 可以表示为如下形式：

$$T_{k,k-1}=\begin{bmatrix}R_{k,k-1}&t_{k,k-1}\\0&1\end{bmatrix} \tag{2}$$

这里：$R_{k,k-1}\in SO(3)$ 是旋转矩阵，$t_{k,k-1}\in\mathbb{R}^{3\times1}$ 是平移向量。

考虑到式（1）中使用的是点 p 在当前帧相机坐标系下的坐标。由于相机在运动，所以 p 在当前帧中的坐标应该是它在前一帧中的坐标根据相机的位姿变换得到的结果，于是有

$$\begin{pmatrix}u\\v\\1\end{pmatrix}=\begin{pmatrix}f_u&0&c_u\\0&f_v&c_v\\0&0&1\end{pmatrix}\begin{bmatrix}R_{k,k-1}&t_{k,k-1}\\0&1\end{bmatrix}\begin{pmatrix}X\\Y\\Z\\1\end{pmatrix}=KT_{k,k\text{-}1}\begin{pmatrix}X\\Y\\Z\\1\end{pmatrix} \tag{3}$$

上式两侧都是齐次坐标，因为齐次坐标乘上非零常数后表达同样的含义，所以这里去掉了 Z 。$T_{k,k\text{-}1}$ 又称为相机的外部参数矩阵，它是 VO 中待估计的目标。$T_{k,k\text{-}1}$ 的计算方法将在本文 2.2.2 节中介绍。

为了便于表达，将 $T_{k,k\text{-}1}$ 的逆矩阵 $T_{k\text{-}1,k}$ 简写为 T_k 。假设集合 $T_{1:n}=\{T_1,\ldots,T_n\}$ 包含所有相邻帧之间的相对运动。集合 $C_{0:n}=\{C_0,\ldots,C_n\}$ 包含所有相对于 $k=0$ 时刻的初始坐标的位姿。当前位姿 $\boldsymbol{C}_n$ 可以由相邻帧之间的相对运动 $T_k(k=1\ldots n)$ 以及 $k=0$ 时刻的初始位姿 $\boldsymbol{C}_0$ 计算得到，即

$$C_n=C_{n-1}T_n=C_0T_1\ldots T_n \tag{4}$$

VO 的主要任务是从图像 I_k 和 I_{k-1} 中计算出相对运动 T_k ，进而恢复出相机的全部轨迹 $C_{0:n}$ 。

1.2 视觉里程计分类

厘清 VO 的分类有助于从宏观上理解整个领域的概况。从不同的角度来看 VO 可以有多种分类方法。按照相机的类型 VO 可以分为单目、立体和 RGB-D 三类；按照利用的图像信息可以分为特征法和直接法的 VO；按照减少漂移的方法可以分为采用滤波器和非线性优化法的 VO。表 1 列出了一些常见的 VO 系统及其分类。

表 1 常见的 VO 系统及其分类

系统名称	相机类型	图像信息	减少漂移
SVO[12]	单目	直接法	非线性优化
PTAM[13]	单目	特征法	非线性优化
ORB-SLAM2[14, 15]	单目、立体、RGB-D	特征法	非线性优化
VISO2[16]	单目、立体	特征法	滤波器
LSD-SLAM[17]	单目	直接法	非线性优化
TLBBA[18]	立体	特征法	非线性优化
MonoSLAM[19]	单目	特征法	滤波器
DEMO[20]	RGB-D	特征法	非线性优化
RTAB-MAP[21]	立体、RGB-D	特征法	非线性优化

1.2.1 单目、立体和 RGB-D

单目视觉 VO 系统使用单目相机作为传感器。单目视觉的优势在于传感器简单、成本低、环境适应性强。但是它的最大问题是无法确定目标的绝对深度，因此所获得的运动轨迹的尺度是模糊的。解决这个问题要求已知空间中某两点之间的距离，或者结合其他的传感器，比如激光雷达、IMU 等。另一方面，单目视觉估计空间点位置需要依靠前后两帧之间的三角测量。而且，相机的运动不能是单纯的旋转。这些问题使单目视觉 VO 的应用受到了一定的限制。

立体视觉 VO 系统使用双目或者多目相机作为传感器。立体视觉利用相机之间固定的基线获取深度信息，能够避免单目视觉的尺度模糊问题，只需要在一帧内即可完成三角测量。但是，立体视觉相机成本较高，相机标定较复杂。测量深度需要进行立体匹配，这个过程非常耗时。而且，其量程受到基线长度的限制，当测量深度远大于基线时，立体视觉就会退化为单目视觉。单目和立体视觉都有各自的优势和不足，目前来看，二者处于均衡的发展状态。

RGB-D 相机兴起于 2010 年左右，它能够同时获得颜色和深度信息。相比于立体视觉，RGB-D 相机能够节省大量计算深度信息的时间。然而，目前大多数 RGB-D 相机存在测量范围小、对日光敏感等问题，因此主要用于室内环境。尽管 RGB-D 相机出现的时间较晚，但发展迅速，有很多优秀的 VO 方案[20, 21]都是基于 RGB-D 相机的。

1.2.2 特征法和直接法

特征法是从稠密图像数据中提取出一些显著特征进行计算。使用特征法的 VO 系统运行稳定、计算成本较低，对光照、图像噪声等不敏感。特征法 VO 的缺点是不适合在缺少特征的场景中应用，如渐变的图像。直接法是利用图像或某个子区域中所有像素的灰度信息计算相机的运动。使用直接法的 VO 系统不要求图像中有特征点，只需要有像素梯度即可。它充分利用了图像信息，有利于实现构建稠密地图等其他的视觉应用。但是直接法的计算量较大，不适合大运动的情况，而且直接法要求图像必须满足像素灰度值不变的假设，而这种假设会由于光照等原因被破坏。根据使用像素数量的不同，直接法可以分为稀疏、半稠密和稠密三种。尽管近年来直接法的 VO 出现了 SVO[12]、LSD-SLAM[17]等一些应用，但是目前成熟的方案较少，主流的 VO 依然采用特征法。

2　视觉里程计关键技术

VO 遵循特征模块、帧间位姿估计和减少漂移的理论框架，实现流程如图 2 所示。特征模块包括特征检测和特征匹配。每获取一帧新的图像，算法首先要检测一些显著性强、可重复性高的图像特征用于位姿估计；然后在当前帧与前一帧图像之间进行特征匹配。特征匹配的目的是在两帧图像中找到特征点对，特征点对是相同的三维空间点在两帧图像上投影产生的二维点。帧间位姿估计包括外点排除和运动估计。通过特征匹配产生的特征点对通常会包含一些不符合数学模型的异常数据，这些数据点被称为外点。外点对于运动估计会产生严重的影响，因此需要排除这些点。接下来是根据余下的特征点对计算当前帧与前一帧之间相机的相对运动，也就是运动估计。外点排除和运动估计通常是一个迭代的过程。由于所获得的两帧间的位姿估计不可避免的会产生误差累积，因此需要使用一些减少漂移的方法获得更精确的相机位姿，主要有滤波器法和非线性优化法。

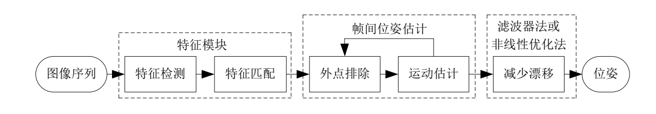


图 2　VO 实现流程

2.1　特征模块

特征模块是之后的位姿估计的基础，下面分别介绍特征的检测和匹配。

2.1.1 特征检测

经典的特征检测算法主要有 Moravec、Forstner、Harris、Shi-Tomasi、SUSAN、FAST、SIFT、SURF、MSER 以及 Censure 等，其中以 Harris[22]和 SIFT[23]两种算法应用最为广泛。Harris 角点对噪声以及旋转变化都具有较强的稳定性，能够提供丰富的信息，是基于视觉的位姿估计系统常用的特征检测算法[5, 20]。然而，Harris 角点对尺度和仿射变换都较为敏感，parra 等[24]证明了 Harris 角点在场景中存在重复的纹理特征时，很容易产生误匹配，他们认为 SIFT 特征点更适合应用于 VO 系统中。SIFT 算法对旋转和尺度具有不变性，对光照、视角变化和噪声也有较强的稳定性，因而很多 VO 系统采用了 SIFT 特征[24, 25]。SIFT 算法的缺点是计算效率很低，为了满足实时性的要求，TLBBA[18]充分考虑了 VO 工作的具体条件，简化了 SIFT 算法，通过引入 GPU 使特征跟踪频率达到 40HZ。

近年来，研究人员在经典的特征检测算法的基础上提出了许多新的算法。2011 年，Rublee 等人在 FAST 和 BRIEF 算法的基础上提出了 ORB 算法[26]，它具有良好的旋转和尺度不变性，速度是 SIFT 的 30-50 倍。ORB 被成功应用于著名的 ORB-SLAM[14, 15]中，表明它是一种兼顾了精度和效率的优秀的特征算法。同年，Leutenegger 等人提出了 BRISK 算法[27]，采用自适应通用加速分割算法，其特征检测速度比 ORB 更快。2012 年，Alcantarilla 等人提出了基于非线性尺度空间理论的 KAZE 算法[28]，与 SIFT 相比，KAZE 具有更好的尺度和旋转不变性。2013 年，他们又发布了改进的 A-KAZE 算法[29]，计算速度有很大提高。

2.1.2 特征匹配

在完成特征检测之后，需要将每个特征点及其邻域转换成一个紧致的描述符，以便与其他的描述符相匹配。经典的特征描述符是 SIFT[23]以及由它发展而来的 SURF[30]。SIFT 已被证明对光照、旋转、尺度和高达 60 度的视角变换都具有很好的稳定性。SURF 用箱型滤波器来近似高斯差分滤波器，能够带来更高的计算效率。此外，还有一些描述符从 SIFT 发展而来，如

PCA-SIFT[31]、DAISY[32]等，这些描述符主要是针对 SIFT 效率低的缺点加以改进。

从 2010 年开始出现了一些二进制字符串描述符，包括 BRIEF[33]、ORB[26]、BRISK[27]、FREAK[34]和 NESTED[35]等，这些描述符的效率普遍高于浮点型描述符。Hartmann 等人[36]比较了 SIFT、SURF、BRIEF、ORB、BRISK 及 FREAK 等六种描述符，结果显示 SIFT 依然获得了最高的精度，如果将计算效率作为所要考虑的主要因素，那么 BRIEF 是最好的选择。Khan 等人[37]使用八种图像数据集比较了当前流行的一些特征描述符，结果显示 SIFT 的综合表现最好。在二进制描述符中，NESTED 获得了最佳的效果。值得注意的是，他们从每种图像数据集中所得到的比较结果不尽相同，这说明特征算法的性能与场景有关。

特征匹配就是要通过一定的相似性度量准则在两组特征点集合中找到对应的特征点对。相似性度量准则包括欧式距离和汉明距离等。通常欧氏距离适合浮点型特征描述符，而汉明距离适合二进制字符串描述符。特征匹配过程中涉及到的一个重要问题是匹配搜索算法。最常用的匹配搜索算法是近似最近邻（ANN)搜索。ANN 能够以较小的准确率的代价换取搜索速度的大幅提升。1997 年，Beis 等人提出了一种基于近似 Kd-树的 BBF 算法[38]。该算法能够以 95%的概率找到最近邻点，并且将速度提高 1000 倍，因此得到广泛应用[23, 24]。其他 ANN 算法还有 Spill-tree[39]、分层 K-均值树[40]和随机 kd-树[41]等。此外，ORB-SLAM[14, 15]采用的词袋模型[42]，LSD-SLAM[17]采用的 FAB-MAP 方法[43]都是高效的匹配搜索算法。为了加快搜索速度，还可以为搜索区域添加一定的约束，如运动模型约束[16, 44]、三维特征点位置约束[45]和极线约束[24]等。

2.2 帧间位姿估计

2.2.1 外点排除

外点的来源主要有两个途径：a)误匹配，产生误匹配的原因包括图像噪声、光照、视角变化以及特征匹配算法本身所产生的误匹配等；b)场景中的运动目标。这些外点会对运动估计产生重大影响，为了获得精确的运动估计结果必须去除它们。

一种经典而有效的排除外点的方法是随机抽样一致性（RANSAC）算法[46]，RANSAC 算法通过迭代可以从含有大量外点的数据集合中提取最优子集。RANSAC 算法的基本思想是从全集中随机抽取一个样本集，计算模型参数，然后用所得到的模型参数验证全集中的其他数据点。经过多次迭代，能够在数据点中获得最高一致性的模型参数就作为模型的解，而那些与这个模型参数不一致的数据点就作为外点。样本集的大小一般设定为能够解出模型参数的最小值，例如在立体视觉中，这个数值通常为 3。RANSAC 算法需要精心选取的主要参数是迭代次数 M，M 可以用如下公式来估计：

$$M = \frac{\log(1-p)}{\log[1-(1-\varepsilon)^s]} \tag{5}$$

其中：s 表示样本集的大小，ε 是估计的全集中某点是外点的概率，p 是要求的获得合理结果的概率。图 3 是一个使用 RANSAC 算法排除外点的例子（红色“+”表示排除掉的点，绿色“+”表示保留的点）。


图 3　RANSAC 算法排除外点

RANSAC 算法是一种非确定性的算法，也就是说每次执行算法所得到的结果都可能是不一样的。它是以一定的概率获得合理的结果，迭代的次数越多，这种概率就越大。实际上，RANSAC 算法已经成为了 VO 系统中排除外点的一种通用方法[44, 47]。近些年出现了一些 RANSAC 算法的改进版本。针对 RANSAC 算法效率低的问题，Pretto 等人提出了 MLESAC 算法[48]，与 RANSAC 计算内点的数量不同，MLESAC 通过将误差表示为一个混合模型用以评估假设的相似性。Chum 等人提出了一种指导抽样过程的 PROSAC 算法[49]，这种算法的应用前提是输入数据的先验信息是已知的，即需要已知哪些数据点更可能是外点。类似地，Rusu 等人[50]提出根据最相似特征直方图抽取数据点。Raguram 等人[51]详细分析比较了与 RANSAC 相关的各种算法，最后提出了一种自适应的实时 RANSAC 算法 ARRSAC。

除了 RANSAC 以及它的众多衍生方法之外，人们还从其他角度提出了一些不同的外点排除方法。VISO2-S[16]采用三角剖分投票法去除外点。一些文献[5, 52]还使用了一种 bucketing 技术，这种技术能够使特征点尽可能均匀的分布在整个图像上，这会提高 VO 的精度。然而，VO 领域对于运动目标产生的外点的研究较少，如 PTAM[13]、ORB-SLAM[14, 15]、LSD-SLAM[17]等系统都假设场景是静止不变的,因而不适合变化较大的场景。针对这种情况，浙江大学研发的 RDSLAM 系统[53]能够在线检测场景的变化,识别出改变的三维点。还有一些文献[54, 55]只检测图像中的地面点。Ci 等人[56]根据车辆运动的平滑性原理提出了一种空间位置约束法去除运动点。然而这些方法无法处理短时间内发生重大变化的场景，因此这方面仍然有很大的发展空间。

2.2.2 运动估计

运动估计是 VO 系统中的核心计算步骤。更准确的说，是计算相机在当前图像 $\boldsymbol{I}_k$ 和前一帧图像 $\boldsymbol{I}_{k-1}$ 之间的变换矩阵 $\boldsymbol{T}_{k,k-1}$。通过串联这些单步的运动，能够完整地恢复相机和智能体的轨迹。假设第 $k-1$ 帧和第 k 帧上有两组对应的特征点，根据特征点对是二维的还是三维的，有三种计算 $\boldsymbol{T}_{k,k-1}$ 的方法，即 3D-3D、3D-2D 和 2D-2D。

3D-3D 是从三维点对中求解运动，通常用于立体视觉中。

$T_{k,k-1}$ 可以通过最小化三维点对之间的欧氏距离来进行估计。即

$$T_{k,k-1}=\arg\min_{T_{k,k-1}}\sum_i \left\|Q_k^i - T_{k,k-1}Q_{k-1}^i\right\|^2 \quad (6)$$

其中：i 表示第 i 个特征点，Q_k 和 $\boldsymbol{Q}_{k-1}$ 分别是第 k 帧和第 k-1 帧上特征点的三维仿射坐标，即 $Q=[X,Y,Z,1]^{\mathrm{T}}$。求解这个问题至少需要三对非共线的三维点。尽管使用更多的点对会增加计算量，但是这会提高运动估计的精度，因此通常所使用的点对远大于 3。求解 3D-3D 问题的方法有奇异值分解法[57]、四元数法[58]和迭代最近点算法（ICP）[4]等。

3D-2D 运动估计是从三维空间点和二维图像点对中求解运动，既可以用于单目视觉也可以用于立体视觉。这种方法与 3D-3D 运动估计相似，只是最小化函数为二维重投影误差，代价函数如下：

$$T_{k,k-1}=\arg\min_{T_{k,k-1}}\sum_i \left\|q_k^i - \pi(T_{k,k-1},Q_{k-1}^i)\right\|^2 \quad (7)$$

其中：q_k 是当前帧 I_k 上的特征点，$\pi(\boldsymbol{T}_{k,k-1},\boldsymbol{Q}_{k-1}^i)$ 是 $\boldsymbol{Q}_{k-1}^i$ 在经过运动变换 $T_{k,k-1}$ 之后在 I_k 上的投影函数。3D-2D 也称为 N 点透视投影（PNP），是目前最常用的一种运动估计方法。Moreno-Noguer 等人[59]给出了 PNP 的多种求解方法。求解 3D-2D 问题至少需要 3 个点对，称为 P3P。有关 P3P 问题的研究开始的较早，迄今为止有超过 10 种解法[60]。对于存在外点情况下的 3D-2D 问题，P3P 是鲁棒的运动估计的标准方法。

2D-2D 方法是从二维图像点对中求解运动参数，一般用于单目视觉中。3D-3D 和 3D-2D 方法只有在三维数据能够获得的情况下才能够实现。但是有时这种条件并不能满足，例如在估计单目摄像机所获得的前两帧图像之间的相对变换时，这两帧上的二维点还没有作三角化测量。在这种情况下，极线约束可以用来估计这种变换，图 4 描述了极线约束的几何关系。极线约束给出了同一个三维点在两个不同视角下的几何约束关系。假设点 $\boldsymbol{q}_k$ 是图像 $\boldsymbol{I}_k$ 上某一个特征点，点 q_{k-1} 是图像 $\boldsymbol{I}_{k-1}$ 上 $\boldsymbol{q}_k$ 的特征点对。点 $\boldsymbol{q}_k$ 和点 $\boldsymbol{q}_{k-1}$ 对应的三维点都为点 $\boldsymbol{Q}$。那么点 q_k、q_{k-1}、Q 以及摄像机中心位于同一平面上。根据这一原则，所有对应的图像点满足如下共面方程：

$$q_k^T F_k q_{k-1}=0 \quad (8)$$

这里 F_k 称为基础矩阵。F_k 包含了摄像机在两帧之间的运动以及摄像机的内部参数。

如果摄像机的内部参数矩阵 K 是已知的，那么共面方程可以写为

$$m_k^T E_k m_{k-1}=0 \quad (9)$$

这里：m_k、m_{k-1} 是对应 q_k、q_{k-1} 的归一化图像坐标，即 $m_k=K^{-1}q_k$，$m_{k-1}=K^{-1}q_{k-1}$，E_k 称为本质矩阵。本质矩阵 $\boldsymbol{E}_k$ 包含了摄像机的旋转和平移运动参数，其中平移运动参数由一个尺度因子决定。

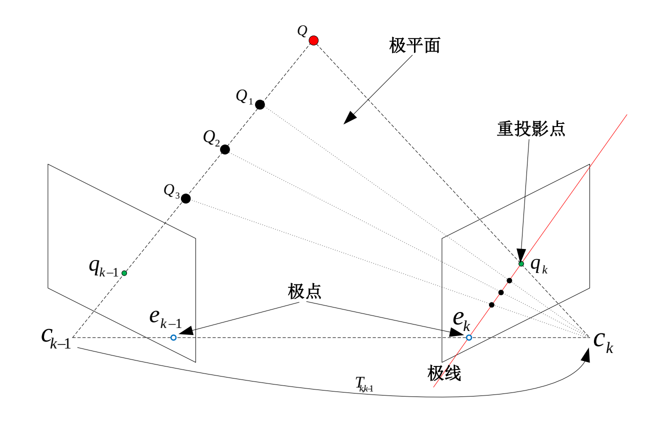
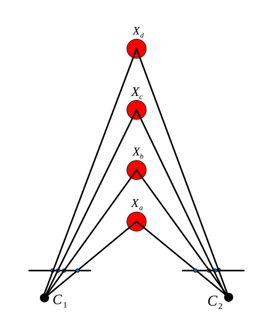


图 4　极线约束的几何描述

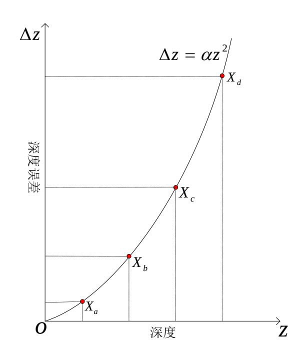
对于已经标定的相机，求解 2D-2D 问题至少需要 5 个二维点对，5 点算法与鲁棒估计器（如 RANSAC 算法）结合能够获得精确的位姿估计结果，特别是 Nister 等人提出的高效 5 点算法[61]及其改进版[62]已经成为存在外点情况下求解 2D-2D 问题的标准方法。此外，求解 2D-2D 问题的方法还有 6 点算法[63]、7 点算法[64]和 8 点算法[65]。Stewenius 等人[62]对 2D-2D 问题的多种求解方法做了比较，结果显示高效 5 点算法的综合性能最佳。有些车用 VO 采用了运动模型约束，运动估计所需的点对数还可以更少，从而减少了计算时间。例如，Fraundorfer 等人[66]提出了相机在已知两个旋转角情况下的 3 点算法。在相机做平面运动时，运动模型的复杂度降为 3 个自由度，此时只需要两个点对。Scaramuzza 等人[67]通过引入车辆运动的非完整性约束，使运动模型的复杂度降为 2 个自由度，他们只利用 1 对特征点求取车辆运动的模型解，其运动估计频率高达 400 Hz。

一般而言，3D-2D 的精度高于 3D-3D。原因是特征点的三角测量在深度方向上有很大的不确定性。尤其是对于距离较远的点，三维点在深度方向上的变化并不会在投影的图像位置上产生很大的变化，如图 5(a)所示。李海滨等人[68]证明了测量深度的误差(Δz)与深度平方(z^2)成正比，图 5(b)描述了二者之间的关系。在使用 3D-3D 时，这种不确定性对于运动估计会造成很严重的影响。而 3D-2D 中的代价函数是图像上的重投影误差，这种不确定性在重投影的过程中在很大程度上被抵消了。此外，为了从根本上解决这种深度上的不确定性问题，Forster 等人[12]在他们的 SVO 系统中提出了深度滤波器的概念，推导了基于均匀-高斯混合分布的深度滤波器。他们将深度滤波器用于特征点三维位置估计，获得了较好的效果。

在单目视觉中，2D-2D 虽然不需要三角测量，但实践中 3D-2D 比 2D-2D 应用更多，原因是 3D-2D 数据关联的速度更快。为了精确的运动估计，外点的排除是一项非常重要的工作，而这种操作所需要的时间与运动估计所必须的最少特征点数目紧密相关。如前所述，2D-2D 需要至少 5 个点对，但是 3D-2D 最少只需要 3 个点对，这使得 3D-2D 方法的运动估计速度更快。因此 2D-2D 通常只用于单目 VO 的初始化。



(a)三维点的深度与其投影点的图像位置的关系



(b) 测量深度的误差(Δz)与深度(z)关系曲线

图 5 三角测量在深度方向上的不确定性

2.3 减少漂移

如本文 1.1 节所述，相机的当前位姿 $\boldsymbol{C}_n$ 是由单步的相对运动 $T_k (k=1...n)$ 以及 $k=0$ 时刻的初始位姿 $\boldsymbol{C}_0$ 计算得到的。而 T_k 与实际的相机相对运动之间不可避免的存在一定的误差，因此当前位姿的误差取决于之前的每一次运动估计的误差。Smith 等[69]的误差繁殖定律证明了当前位姿 C_n 的误差会随着单步的相对运动 T_k 的串联而增大。这种误差逐渐增大的现象叫做漂移，图 6 描述了漂移的产生过程。要减少漂移，除了减少帧间相对运动误差之外，主要有滤波器法和非线性优化法两种。

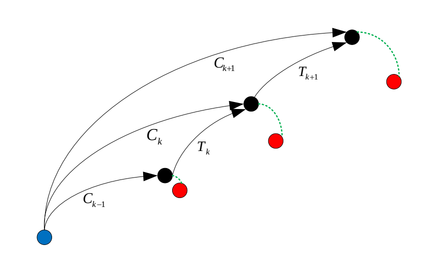


图 6 VO 的漂移现象

（蓝点表示初始位置，红点表示真实位置，黑点表示测量位置）

2.3.1 滤波器法

滤波器法在早期的 VO 中占据主导地位，其中最常用的是扩展卡尔曼滤波器（EKF）[19]。EKF 以相机的当前位姿和所有点的三维坐标为状态变量，更新其均值和协方差。对于 VO 这种非线性系统，EKF 实际上给出了单次线性近似下的最大后验估计。Kitt 等人[52]使用无迹卡尔曼滤波（UKF）获得了比 EKF 更精确的估计结果。与采用一阶泰勒展开的 EKF 相比，UKF 接近 3 阶精度。还有很多算法相对 EKF 能够降低计算的复杂度，如稀疏扩展信息滤波器（SEIF）[70]、Atlas 框架[71]、分治法[72]等。但是滤波器方法有很大的局限性，它在一定程度上假设了马尔可夫性，即 k 时刻的状态只与 $k-1$ 时刻相关，而与之前的状态和观测都无关，依然容易造成误差累积。因此滤波器方法一般用在计算资源受限或待估计量比较简单的场合，而非线性优化是目前的主流方法。

2.3.2 非线性优化法

非线性优化法能够考虑 k 时刻的状态与之前所有状态的关系。PTAM[13]是第一个使用非线性优化方案的系统，它引入了关键帧机制，即不精细的处理每一帧图像，而是把几个关键帧串起来，然后优化其轨迹。在非线性优化方法中，光束法平差（BA）是应用最广泛的一种方法。BA 通过最小化多帧的重投影误差从而优化相机的位姿以及点的三维坐标。当代价函数把所有帧都考虑进来时，BA 叫做全局光束法平差（GBA）。当代价函数把固定的 m 帧考虑进来时，BA 叫做局部光束法平差（LBA）。由于将全部帧考虑在内，GBA 相对于 LBA 优化精度更高，但是 GBA 的计算效率很低。LBA 更适合应用到实时系统中，例如在 ORB-SLAM[14]中就包含了一个 LBA 的优化线程，负责求解更精细的相机位姿和空间点的三维坐标。TLBBA[18]使用了一种两阶段局部双目光束法平差的方法，充分利用双目图像序列中的信息和约束，对运动估计结果进行优化。事实上，目前许多先进的 VO 系统都采用了 LBA 方法[73, 74]。

3 基于深度学习的视觉里程计

以上所述 VO 系统都是利用几何原理恢复相机的运动。不同于上述思路，近几年来，有学者尝试使用深度学习（deep learning，DL）的方法揭示图像光流和相机运动的关系，为 VO 提供了新的解决方案。Konda 等人[75]最先通过提取视觉运动和深度信息实现了基于 DL 的 VO。在使用立体图像估计出深度信息之后，卷积神经网络（convolutional neural network，CNN）通过 softmax 函数预测相机速度和方向的改变。Kendall 等人[76]利用 CNN 实现了输入为 RGB 图像，输出为相机位姿的端到端定位系统。该系统提出了 23 层深度卷积网络 PoseNet，利用迁移学习将分类问题的数据库用于解决复杂的图像回归问题。其训练得到的特征相较于传统局部视觉特征，对于光照、运动模糊以及相机内参等具有更强的鲁棒性。他们提出利用 SfM 自动生成训练样本的标注，不需要人工标注每一幅图像的位姿信息，但是这种方法对于大规模场景非常耗时。由于系统将一个训练好的 CNN 模型作为场景的表观地图，因此在遇到新环境时需要重新训练或者微调，这也是将 DL 用于 VO 时面对的一个最大的问题。为了解决这一问题，Costante 等人[77]用稠密光流代替 RGB 图像作为 CNN 的输入。该系统设计了三种不同的 CNN 架构用于 VO 的特征学习，实现了算法在图像模糊和曝光不足等条件下的鲁棒性。然而，实验结果也表明训练数据对于算法影响很大，当图像序列帧间运动较大时，算法误差很大，这主要是由于训练数据缺少高速训练样本。

Wang 等人[78]使用深度递归卷积神经网络（recurrent convolutional neural networks，RCNN），提出了一种新的端到端

的单目 VO 框架。基于 RCNN，不仅可以使用卷积神经网络自动的为 VO 问题提供有效的特征表示，也可以使用递归神经网络隐式的对运动模型和数据关联模型进行建模。他们在 KITTI 的 VO 数据集上的实验表明，其算法的性能可以和目前最先进的 VO 方法相媲美。但是他们也强调基于 DL 的 VO 并不能取代传统的基于几何方法的 VO，而是一种可行的补充。最新的文献[79]构建了一个自编码深度网络学习产生光流的非线性潜在空间描述。这个自编码网络与另外一个神经网络联合训练以产生相机的自运动估计，两个网络互相借鉴可以获得更加鲁棒的光流描述和更精确的运动估计。

与传统的基于几何方法的 VO 相比，基于 DL 的 VO 无须建立复杂的物体运动的几何模型，甚至无须考虑相机的校准参数以及相对尺度问题，运动估计的准确性和鲁棒性依赖于神经网络估计器的设计和用于训练的图像库是否涵盖待测场景的变化。目前，基于 DL 的 VO 研究仍然处于起步阶段，当测试场景与训练场景存在较大差异时，其性能不够理想。从现有的基于 DL 的 VO 方法来看，不同的实现方法之间神经网络架构存在较大的差异，对各类场景的鲁棒性即网络的泛化能力有待进一步提高。

4　算法评价

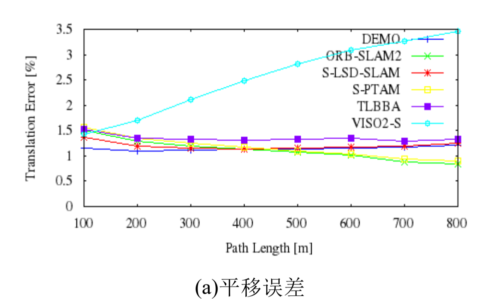
要比较各种 VO 系统的性能，需要在相同的图像序列上进行测试，为此，一些机构提供了公共数据集。本部分介绍目前比较流行的三种数据集：KITTI[80]、Tsukuba[81]和 TUM[82]数据集。表 2 给出了这三种数据集的基本信息。

表 2　三种 VO 数据集的基本信息

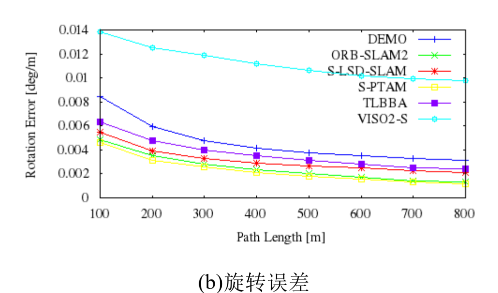
数据集	相机类型	场景	网址
KITTI	立体	室外	http://www.cvlibs.net/datasets/kitti/eval_odometry.php
Tsukuba	立体	人工图像	http://www.cvlab.cs.tsukuba.ac.jp/dataset/tsukubastereo.php
TUM	RGB-D	室内	http://vision.in.tum.de/data/datasets/rgbd-dataset

KITTI 是德国卡尔斯鲁厄理工学院和芝加哥丰田技术研究所联合创办的一个算法评测平台。KITTI 中的图像序列是由行驶的汽车在城市和自然环境中拍摄的，车速、光照以及行驶轨迹的类型多种多样。KITTI 的 VO 模块包括 22 个立体图像序列，其中序列 00 到 10 为训练集，KITTI 为这 11 个序列提供了真实值。图像序列 11 到 21 为测试集，对于这 11 个序列，KITTI 不公开真实值，测试集用于评价各种 VO 系统的性能。此外，KITTI 的设计者还给出了 VO 算法的评价准则，即平均平移误差（ATE）和平均旋转误差（ARE）。图 7 显示了几个流行的 VO 系统随着路径长度增加的旋转误差和平移误差。表 3 列出了它们的评价结果，即 ATE 和 ARE。

Tsukuba 数据集是一个经过真实感渲染的人工图像数据集。这个数据集是一个 1 分钟长的视频，包含了 1800 对带有真实的视差图和遮挡图的立体图像。同时包含了每一帧相机的 3 维位置和方向信息，因此这个数据集也可以用于评价相机跟踪方法。相比于依靠 GPS/IMU 等传感器获取真实值的真实图像数据集，人工图像数据集的主要特点是其真实值不含噪声。



(a)平移误差



(b)旋转误差

图 7　几种 VO 系统随着路径长度增加的平移和旋转误差

表 3　几种 VO 系统在 KITTI 上的评价结果

系统名称	ATE	ARE
DEMO[20]	1.14%	0.0049[deg/m]
ORB-SLAM2[14, 15]	1.15%	0.0027[deg/m]
S-PTAM[83]	1.19%	0.0025[deg/m]
S-LSD-SLAM[17]	1.20%	0.0033[deg/m]
TLBBA[18]	1.36%	0.0038[deg/m]
VISO2-S[16]	2.44%	0.0114[deg/m]

德国慕尼黑工业大学（TUM）提供了针对 RGB-D 和单目等相机的多种数据集，其中最常用的是它的 RGB-D 数据集。该数据集中包含了 39 个在室内环境下拍摄的图像序列，涵盖了各种各样的场景以及相机运动方式。大多数序列由手持的 Kinect 相机拍摄得到，做无约束的 6 自由度运动，还有一部分序列的 Kinect 相机装在 3 个机器人上。该数据集还根据结构和纹理、含有动态目标等特征将场景分类。TUM 数据集提供了真实值，它是由一个外部的运动捕捉装置测量得到的，而且还提供了测评工具。TUM 也给出了算法的评价准则，即相对位姿误差（RPE）和绝对轨迹误差（ATE）。

5　结束语

本文结合一些先进的 VO 系统，对 VO 技术进行了综述。VO 使用摄像机替代传统的传感器，造价较低。它无须场景和运动的先验信息，不存在编码器读数不准和传感器精度降低等因素引起的数据误差，不受不平坦地带车轮打滑以及其他不宜条件的影响。目前，VO 技术已成功应用于陆地、航空和水下等各种移动机器人系统中，此外，VO 还广泛应用于汽车和消费类电子产品等工业中。尽管如此，VO 系统仍然面临一些限制，

如图像缺少纹理，相机快速移动而引起的图像模糊，光照和成像条件的影响，这些都会导致对相机位姿的估计不准确。最大的挑战是如何在室外环境下的长距离运行中保持稳定性。从 VO 面临的这些问题以及近期出现的文献来看，如下研究方向可能是今后研究的趋势：

a）相机的种类多样化。除了单目、立体和 RGB-D 相机之外，还有一些 VO 系统使用了其他类型的相机，如全方位相机[84]、鱼眼相机[85]和反射折射相机[86]等。这些相机往往适合不同的场景，如 Zhang 等[87]经过实验认为大视场角相机适合室内等空间较小的环境，而小视场角相机适合室外等大范围的环境。因此，可以从场景和相机的关系出发，选择合适的相机类型，增强 VO 的环境适应性。

b）设计高性能的特征检测和描述符。近几年，特征算法领域快速发展，新的特征算法不断涌现，如二进制描述符 NESTED[35]对外点排除显示了很好的效果，而 Desai 等[88]提出的 SYBA 特征描述符能够有效减少漂移。另外，利用边缘等更高层的图像信息能够减少算法对于特征的依赖，实际上，LSD-SLAM 就利用了图像的边缘信息。未来一个可能的研究思路是将点特征和边缘特征相结合，从而使 VO 能够更好的应对低纹理的场景。

c）利用已有的基于视觉的运动目标检测算法的成果。文中提到运动目标是外点的一个重要来源，如何去除运动目标点是 VO 研究的一项重要内容。目前，基于视觉的运动目标检测已经取得了大量的研究成果，一个可行的思路是将这些研究成果和 VO 相结合去除运动点，排除运动目标的干扰。这对于提高特征点集的质量，进而提高运动估计的精度具有重要意义。

d）基于 DL 的 VO 无须建立复杂的物体运动的几何模型，可以从图像序列的变化中训练出特征，并且映射出位姿参数，是 VO 研究的一个新方向。如何提高现有的基于 DL 的 VO 对各类场景的鲁棒性是一个关键问题。

参考文献：

[1] Scaramuzza D, Fraundorfer F. Visual odometry: part I: the first 30 years and fundamentals [J]. IEEE Robotics & Automation Magazine, 2011, 18 (4): 80-92.

[2] Moravec H P. Obstacle avoidance and navigation in the real world by a seeing robot rover [D]. Palo Alto: Stanford University, 1980.

[3] Matthies L, Shafer S. Error modeling in stereo navigation [J]. IEEE Journal on Robotics and Automation, 1987, 3 (3): 239-248.

[4] Milella A, Siegwart R. Stereo-based ego-motion estimation using pixel tracking and iterative closest point [C]// Proc of the 4th IEEE International Conference on Computer Vision Systems. 2006: 21-21.

[5] Nister D, Naroditsky O, Bergen J, *et al.* Visual odometry [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2004: 652-659.

[6] Patra S, Aggarwal H, Arora H, *et al.* Computing Egomotion with Local Loop Closures for Egocentric Videos [C]// Proc of IEEE Winter Conference on Applications of Computer Vision. 2017: 454-463.

[7] Wang S, Zhang Y, Zhu F. Monocular visual SLAM algorithm for autonomous vessel sailing in harbor area [C]// Proc of the 25th Saint Petersburg International Conference on Integrated Navigation Systems. 2018: 1-7.

[8] He Hong, Jia Yunhui, Sun Lei. Simultaneous location and map construction based on RBPF-SLAM algorithm [C]// Proc of Chinese Control And Decision Conference. 2018: 4907-4910.

[9] 江燕华，熊光明，姜岩，等. 智能车辆视觉里程计算法研究进展 [J]. 兵工学报, 2012, 33 (2): 214-220. (Jiang Yanhua, Xiong Guangming, Jiang Yan, *et al.* A review of visual odometry for intelligent vehicle [J]. Acta Armamentarii, 2012, 33 (2): 214-220.)

[10] 李宇波，朱效洲，卢惠民，等. 视觉里程计技术综述 [J]. 计算机应用研究, 2012, 29 (8): 2801-2805, 2810. (Li Yubo, Zhu Xiaozhou, Lu Huimin, *et al.* Review on visual odometry technology [J]. Application Research of Computers, 2012, 29 (8): 2801-2805, 2810.)

[11] Fraundorfer F, Scaramuzza D. Visual odometry, part II: matching, robustness, optimization, and applications [J]. IEEE Robotics & Automation Magazine, 2012, 19 (2): 78-90.

[12] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry [C]// Proc of IEEE International Conference on Robotics and Automation. 2014: 15-22.

[13] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces [C]// Proc of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. 2007: 225-234.

[14] Mur-Artal R, Montiel JMM, Tard JD. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. IEEE Trans on Robotics, 2015, 31 (5): 1147-1163.

[15] Mur-Artal R, Tardós JD. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras [J]. IEEE Trans on Robotics, 2017, 33 (5): 1255-1262.

[16] Geiger A, Ziegler J, Stiller C. StereoScan: dense 3D reconstruction in real-time [C]// Proc of IEEE Intelligent Vehicles Symposium. 2011: 963-968.

[17] Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM [C]// Proc of the 13th European Conference on Computer Vision. 2014: 834-849.

[18] Lu Wei, Xiang Zhiyu, Liu Jilin. High-performance visual odometry with two-stage local binocular BA and GPU [C]// Proc of IEEE Intelligent Vehicles Symposium. 2013: 1107-1112.

[19] Davison A J, Reid I D, Molton N D, *et al.* MonoSLAM: real-time single camera SLAM [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29 (6): 1052-1067.

[20] Zhang J, Kaess M, Singh S. Real-time depth enhanced monocular odometry [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2014: 4973-4980.

[21] Labbe M, Michaud F. Online global loop closure detection for large-scale

multi-session graph-based SLAM [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2014: 2661-2666.

[22] Harris C G, Stephens M J. A combined corner and edge detector [C]// Proc of the 4th Alvey Vision Conference. 1988: 147-151.

[23] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60 (2): 91-110.

[24] Parra I, Sotelo M A, Llorca D F, *et al.* Robust visual odometry for vehicle localization in urban environments [J]. Robotica, 2010, 28 (3): 441-452.

[25] Tardif J P, Pavlidis Y, Daniilidis K. Monocular visual odometry in urban environments using an omnidirectional camera [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2008: 2531-2538.

[26] Rublee E, Rabaud V, Konolige K, *et al.* ORB: an efficient alternative to SIFT or SURF [C]// Proc of International Conference on Computer Vision. 2011: 2564-2571.

[27] Leutenegger S, Chli M, Siegwart RY. BRISK: binary robust invariant scalable keypoints [C]// Proc of International Conference on Computer Vision. 2011: 2548-2555.

[28] Alcantarilla P F, Bartoli A, Davison A J. KAZE features [C]// Proc of the 12th European Conference on Computer Vision. 2012: 214-227.

[29] Alcantarilla P F, Nuevo J, Bartoli A. Fast explicit diffusion for accelerated features in nonlinear scale spaces [C]// Proc of International Conference on Control, Automation and Systems. 2013: 704-709.

[30] Bay H, Tuytelaars T, Van Gool L. SURF: speeded up robust features [C]// Proc of the 9th European Conference on Computer Vision. 2006: 404-417.

[31] Yan K, Sukthankar R. PCA-SIFT: a more distinctive representation for local image descriptors [C]// Proc ofIEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004: 506-513.

[32] Tola E, Lepetit V, Fua P. DAISY: an efficient dense descriptor applied to wide-baseline stereo [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32 (5): 815-830.

[33] Calonder M, Lepetit V, Strecha C, *et al.* BRIEF: binary robust independent elementary features [C]// Proc of the 11th European Conference on Computer Vision. 2010: 778-792.

[34] Vandergheynst P, Ortiz R, Alahi A. FREAK: fast retina keypoint [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 510-517.

[35] Byrne J, Shi J. Nested shape descriptors [C]// Proc of IEEE International Conference on Computer Vision. 2013: 1201-1208.

[36] Hartmann J, Klussendorff JH, Maehle E. A comparison of feature descriptors for visual SLAM [C]// Proc of European Conference on Mobile Robots. 2013: 56-61.

[37] Khan N, Mccane B, Mills S. Better than SIFT? [J]. Machine Vision and Applications, 2015, 26 (6): 819-836.

[38] Beis J S, Lowe D G. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1997: 1000-1006.

[39] Liu T, Moore A W, Gray A, *et al.* An investigation of practical approximate nearest neighbor algorithms [C]// Proc of International Conference on Neural Information Processing Systems. 2004: 825-832.

[40] Nister D, Stewenius H. Scalable recognition with a vocabulary tree [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006: 2161-2168.

[41] Silpa-Anan C, Hartley R. Optimised KD-trees for fast image descriptor matching [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-8.

[42] Galvez-Lopez D, Tardos J D. Bags of binary words for fast place recognition in image sequences [J]. IEEE Trans on Robotics, 2012, 28 (5): 1188-1197.

[43] Glover A, Maddern W, Warren M, *et al.* OpenFABMAP: an open source toolbox for appearance-based loop closure detection [C]// Proc of IEEE International Conference on Robotics and Automation. 2012: 4730-4735.

[44] Maimone M, Cheng Y, Matthies L. Two years of visual odometry on the mars exploration rovers [J]. Journal of Field Robotics, 2007, 24 (3): 169-186.

[45] Davison A J. Real-time simultaneous localisation and mapping with a single camera [C]// Proc of the 9th IEEE International Conference on Computer Vision. 2003: 1403-1410.

[46] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [M]// Readings in Computer Vision. San Francisco: Morgan Kaufmann, 1987: 726-740.

[47] Deigmoeller J, Eggert J. Stereo visual odometry without temporal filtering [C]// Proc of the 38th German Conference on Pattern Recognition. 2016: 166-175.

[48] Pretto A, Menegatti E, Bennewitz M, *et al.* A visual odometry framework robust to motion blur [C]// Proc of IEEE International Conference on Robotics and Automation. 2009: 2250-2257.

[49] Chum O, Matas J. Matching with PROSAC-progressive sample consensus [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005: 220-226.

[50] Rusu RB, Blodow N, Beetz M. Fast point feature histograms (FPFH) for 3D registration [C]// Proc of IEEE International Conference on Robotics and Automation. 2009: 3212-3217.

[51] Raguram R, Frahm J M, Pollefeys M. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus [C]// Proc of the 10th European Conference on Computer Vision. 2008: 500-513.

[52] Kitt B, Geiger A, Lategahn H. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme [C]// Proc of IEEE Intelligent Vehicles Symposium. 2010: 486-492.

[53] Tan Wei, Liu Haomin, Dong Zilong, *et al.* Robust monocular SLAM in dynamic environments [C]// Proc of IEEE International Symposium on Mixed and Augmented Reality. 2013: 209-218.

[54] Musleh B, Martin D, Escalera Adl, *et al.* Visual ego motion estimation in urban environments based on U-V disparity [C]// Proc of IEEE Intelligent Vehicles Symposium. 2012: 444-449.

[55] Min Qin, Huang Yingping. Motion detection using binocular image flow in dynamic scenes [J]. EURASIP Journal on Advances in Signal Processing, 2016, 2016 (1): 1-12.

[56] Ci Wenyan, Huang Yingping. A robust method for ego-motion estimation in urban environment using stereo camera [J]. Sensors (Basel) , 2016, 16 (10): 1-14.

[57] Gong Piliang, Zhang Qifeng, Zhang Aiqun. Stereo vision based motion estimation for underwater vehicles [C]// Proc of the 2nd International Conference on Intelligent Computation Technology and Automation. 2009: 745-749.

[58] Horn B K P. Closed-form solution of absolute orientation using unit quaternions [J]. Journal of the Optical Society of America A, 1987, 4 (4): 629-642.

[59] Moreno-Noguer F, Lepetit V, Fua P. Accurate non-iterative o (n) solution to the PnP problem [C]// Proc of the 11th IEEE International Conference on Computer Vision. 2007: 1-8.

[60] Haralick B M, Lee C N, Ottenberg K, *et al.* Review and analysis of solutions of the three point perspective pose estimation problem [J]. International Journal of Computer Vision, 1994, 13 (3): 331-356.

[61] Nister D. An efficient solution to the five-point relative pose problem [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26 (6): 756-770.

[62] Stewenius H, Engels C, Nister D. Recent developments on direct relative orientation [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2006, 60 (4): 284-294.

[63] Pizarro O, Eustice R, Singh H. Relative pose estimation for instrumented, calibrated imaging platforms [C]// Proc of Digital Image Computing Techniques & Applications. 2003: 601-612.

[64] Hartley R, Zisserman A. Multiple view geometry in computer vision [M]// Cambridge: Cambridge University Press, 2003: 1865-1872.

[65] Mirabdollah M H, Mertsching B. Single camera motion estimation: modification of the 8-point method [C]// Proc of the 6th International Conference on Intelligent Robotics and Applications. 2013: 117-128.

[66] Fraundorfer F, Tanskanen P, Pollefeys M. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles [C]// Proc of the 11th European Conference on Computer Vision. 2010: 269-282.

[67] Scaramuzza D. 1-point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints [J]. International Journal of Computer Vision, 2011, 95 (1): 74-85.

[68] 李海滨，单文军，刘彬．双目立体视觉测距系统误差模型的研究 [J]. 光学技术, 2006, 32 (1): 24-26. (Li Haibin, Shan Wenjun, Liu Bin. Research of error-model on two eyes stereoscopic measurement system [J]. Optical Technique, 2006, 32 (1): 24-26.)

[69] Smith RC, Cheeseman P. On the representation and estimation of spatial uncertainly [J]. International Journal of Robotics Research, 1986, 5 (4): 56-68.

[70] Thrun S, Koller D, Ghahramani Z, *et al.* Simultaneous Mapping and Localization with Sparse Extended Information Filters: Theory and Initial Results [J]. Springer Tracts in Advanced Robotics, 2004, 7 (1): 363-380.

[71] Bosse M, Newman P, Leonard J, *et al.* An Atlas framework for scalable mapping [C]// Proc of IEEE International Conference on Robotics and Automation. 2003: 1899-1906.

[72] Paz L M, PiniÉs P, TardÓs J D, *et al.* Large-Scale 6-DOF SLAM With Stereo-in-Hand [J]. IEEE Trans on Robotics, 2008, 24 (5): 946-957.

[73] Engel J, Koltun V, Cremers D. Direct sparse odometry [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017: 1-14.

[74] Leutenegger S, Lynen S, Bosse M, *et al.* Keyframe-based visual-inertial odometry using nonlinear optimization [J]. International Journal of Robotics Research, 2015, 34 (3): 314-334.

[75] Konda K, Memisevic R. Learning Visual Odometry with a Convolutional Network [C]// Proc of International Conference on Computer Vision Theory and Applications. 2015: 486-490.

[76] Kendall A, Grimes M, Cipolla R. Convolutional networks for real-time 6-DOF camera relocalization [J]. Education for Information, 2015, 31: 125-141.

[77] Costante G, Mancini M, Valigi P, *et al.* Exploring representation learning with cnns for frame-to-frame ego-motion estimation [J]. IEEE Robotics and Automation Letters, 2016, 1 (1): 18-25.

[78] Wang S, Clark R, Wen H, *et al.* DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks [C]// Proc of IEEE International Conference on Robotics and Automation. 2017: 2043-2050.

[79] Costante G, Ciarfuglia TA. LS-VO: learning dense optical subspace for robust visual odometry estimation [J]. IEEE Robostics and Automation Letters, 2018, 3 (3): 1735-1742.

[80] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving?The KITTI vision benchmark suite [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 3354-3361.

[81] Peris M, Martull S, Maki A, *et al.* Towards a simulation driven stereo vision system [C]// Proc of the 21st International Conference on Pattern Recognition. 2012: 1038-1042.

[82] Sturm J, Engelhard N, Endres F, *et al.* A benchmark for the evaluation of RGB-D SLAM systems [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2012: 573-580.

[83] Pire T, Fischer T, Civera J, *et al.* Stereo parallel tracking and mapping for robot localization [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems (IROS) . 2015: 1373-1378.

[84] Reich S, Seer M, Berscheid L, *et al.* Omnidirectional Visual Odometry for

Flying Robots using Low-power Hardware [C]// Proc of International Conference on Computer Vision Theory and Applications. 2018: 499-507.

[85] Matsuki H, Stumberg Lv, Usenko V, *et al.* Omnidirectional DSO: direct sparse odometry with fisheye cameras [J]. IEEE Robotics and Automation Letters, 2018, 3 (4): 3693-3700.

[86] Ilizirov G, Filin S. Pose Estimation and Mapping Using Catadioptric Cameras with Spherical Mirrors [J]. International Archives of the Photogrammetry Remote Sensing & S, 2016, XLI-B3: 43-47.

[87] Zhang Z, Rebecq H, Forster C, *et al.* Benefit of large field-of-view cameras for visual odometry [C]// Proc of IEEE International Conference on Robotics and Automation. 2016: 801-808.

[88] Desai A, Lee DJ. Visual odometry drift reduction using SYBA descriptor and feature transformation [J]. IEEE Trans on Intelligent Transportation Systems, 2016, 17 (7): 1839-1851.